WPS5550

Policy Research Working Paper 5550

Using Repeated Cross-Sections to Explore Movements in and out of Poverty

Hai-Anh Dang
Peter Lanjouw
Jill Luoto
David McKenzie

The World Bank
Development Research Group
Poverty and Inequality Team
and Finance and Private Sector Development Team
January 2011

Policy Research Working Paper 5550

Abstract

Movements in and out of poverty are of core interest to both policymakers and economists. Yet the panel data needed to analyze such movements are rare. In this paper, the authors build on the methodology used to construct poverty maps to show how repeated cross-sections of household survey data can allow inferences to be made about movements in and out of poverty. They illustrate that the method permits the estimation of *bounds* on mobility, and provide non-parametric and parametric approaches to obtaining these bounds. They test how well the method works on data sets for Vietnam and Indonesia where we are able to compare our method to true panel estimates. The results are sufficiently encouraging to offer the prospect of some limited, basic, insights into mobility and poverty duration in settings where historically it was judged that the data necessary for such analysis were unavailable.

This paper is a product of the Poverty and Inequality Team, and the Finance and Private Sector Development Team; Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at planjouw@worldbank.org and dmckenzie@worldbank.org.

Using Repeated Cross-Sections to Explore Movements into and out of Poverty*

Hai-Anh Dang, World Bank
Peter Lanjouw, World Bank
Jill Luoto, RAND Corporation
David McKenzie, World Bank, BREAD, CEPR and IZA

Keywords: Transitory and Chronic poverty; Synthetic panels; Mobility.

JEL Codes: O15, I32.

* We are grateful to the editor, three anonymous referees, Chris Elbers, Roy van der Weide, and seminar participants at Cornell, Georgetown, Minnesota, and the World Bank for useful comments. This paper represents the views of the authors only and should not be taken to reflect those of the World Bank or any affiliated organization.

"But the whole picture of poverty is not contained in a snapshot income-distribution decile graph. It says nothing about the vital concept of mobility: the potential for people to get out of a lower decile – and the speed at which they can do so."

UK Prime Minister David Cameron, October 2010[1]

1. Introduction

Income mobility is currently at the forefront of policy debates around the world. The prolonged global recession has thrust renewed attention on the problem of chronic poverty, while discussion of widening inequality (particularly driven by high incomes of the top 1%) has led to debate about the extent to which opportunities to succeed are open to all.[2] Policies to address poverty will likely differ depending on whether poverty is transitory (in which case safety net policies will likely be the focus) or chronic (in which case more activist policies designed to remove poverty traps may be designed). However, despite the importance of mobility for policy, in many countries, especially developing countries, there is a paucity of evidence on the duration of poverty and on income mobility due to a lack of panel data.

To overcome the non-availability of panel data, there have been a number of studies, starting with Deaton (1985), that develop pseudo-panels out of multiple rounds of cross-sectional data. Compared to analysis using cross sections, pseudo-panels constructed on the basis of age cohorts followed across multiple surveys have permitted rich investigations into the dynamics of income and consumption over time (e.g., Deaton and Paxson , 1994; Banks, Blundell, and Brugiavini, 2001; and Pencavel, 2007) and of cohort-level mobility (Antman and McKenzie, 2007). However, some of these methods rely on having many rounds of repeated cross-sections (Bourguignon et al, 2004), and the use of cohort-means precludes the examination of income mobility at a level more disaggregated than that of the cohort. As a result, such methods may be of limited appeal to policy makers interested in the mobility of certain (disadvantaged) population groups, or to economists concerned with mobility due to idiosyncratic shocks to income or consumption.

[1] Taken from a commentary "What you receive should depend on how you behave" in *The Independent*, October 10, 2010, http://www.independent.co.uk/opinion/commentators/david-cameron-what-you-receive-should-depend-on-how-you-behave-2102576.html

[2] In the U.S., for example, Alan Krueger's January 2012 address to the Center for American Progress focused heavily on income mobility and was followed by substantial discussion in both national media and in economics blogs. See http://www.whitehouse.gov/sites/default/files/krueger_cap_speech_final_remarks.pdf for the speech.

The purpose of this paper is to introduce and explore an alternative statistical methodology for analyzing movements in and out of poverty based on two or more rounds of cross-sectional data. The method is less data-demanding than many traditional pseudo-panel studies, and importantly allows for investigation of income mobility within as well as between cohorts.[3] The approach builds on an "out-of-sample" imputation methodology described in Elbers et al (2003) for small-area estimation of poverty (the development of "poverty maps"). A model of consumption (or income) is estimated in the first round of cross-section data, using a specification which includes only time-invariant covariates. Parameter estimates from this model are then applied to the same time-invariant regressors in the second survey round to provide an estimate of the (unobserved) first period's consumption or income for the individuals surveyed in that second round. Analysis of mobility can then be based on the actual consumption observed in the second round along with this estimate from the first round.

Although exact point estimates of poverty transitions and income mobility require knowledge of the underlying autocorrelation structure of the income or consumption generating process, we show that, under mild assumptions, one can derive upper and lower bounds on entry into and exit from poverty. We provide two approaches to estimating these bounds. The first is a non-parametric approach, which imposes no structure on the underlying error distribution. We show that the width of the bounds provided by this approach depends on the extent to which time-invariant and deterministic characteristics explain cross-sectional income or consumption. However, in many cases, while the exact autocorrelation is unknown, evidence from other data sources might be available, suggesting that the true autocorrelation lies within a much narrower (and known) range than the extreme values of zero and one underpinning the non-parametric bounds. We provide a parametric bounding approach that can be used in such cases, which imposes more assumptions but permits a narrowing of the bounds relative to the non-parametric case.

[3] Güell and Hu (2006) provide a GMM estimator for the probability of exiting unemployment that also permits disaggregation to the individual level using multiple cross-sections. However, Guell and Hu's method is most appropriate for duration analysis and can only be applied to two rounds of cross sections given two additional conditions: i) availability of data on the duration of unemployment spells, and ii) the two cross sections must have the same population mean and be independent of each other. In this paper our focus is on poverty mobility, and we require simpler data and much less restrictive assumptions to derive lower and upper bounds on poverty mobility. See also Gibson (2001) for a somewhat related literature on how panel data on a subset of individuals can be used to infer chronic poverty for a larger sample, and Foster (2009) and Hojman and Kast (2009) for recent studies that investigate poverty mobility using actual panel data.

To illustrate our methods and examine their performance in practice, we implement both the non-parametric and the parametric bounding methods in two empirical settings: Vietnam and Indonesia. Genuine panel data are available in these settings, and this allows us to validate our method by sampling repeated cross-sections from the panel, constructing mobility estimates using these cross-sections, and then comparing the results to those obtained using the actual panel data. We find that the "true" estimate of the extent of mobility (as revealed by the actual panel data) is generally sandwiched between our upper-bound and lower-bound assessments of mobility. Our analysis reveals further that the width between the upper- and lower-bound estimates of mobility is narrowed as the prediction models are more richly specified, as well as with the addition of the parametric assumption. We thus believe our method may be readily employed to study mobility for a wide variety of situations where only repeated cross sections are available.

The remainder of the paper is structured as follows: Section 2 provides a theoretical framework for obtaining upper and lower bounds on movements into and out of poverty. Sections 3 and 4 describe our non-parametric and parametric estimation methods respectively. Section 5 examines robustness to the choice of poverty line and provides an application to mobility profiling. Section 6 concludes.

2. Theoretical Bounds for Movements In and Out of Poverty with Repeated Cross-Sections

For ease of exposition we consider the case of two rounds of cross-sectional surveys, denoted round 1 and round 2. We assume that both survey rounds are random samples of the underlying population of interest, and each consist of a sample of N_1 and N_2 households respectively.

Let x_{i1} be a vector of characteristics of household i in survey round 1 which are observed (for different households) in both the round 1 and round 2 surveys. This will include such time-invariant characteristics as language, religion, and ethnicity, and if the identity of the household head remains constant across rounds, will also include time-invariant characteristics of the household head such as sex, education, place of birth, and parental education as well as deterministic characteristics such as age. Importantly, x_{i1} can also include time-varying

characteristics of the household that can be easily recalled for round 1 in round 2. Thus variables such as whether or not the household head is employed in round 1, and his or her occupation, as well as their place of residence in round 1 could be included in x_{i1} if asked in round 2.[4]

Then for the population as a whole, the linear projection of round 1 consumption or income, y_{i1}, onto x_{i1} is given by:

$$y_{i1} = \beta_1' x_{i1} + \varepsilon_{i1} \qquad (1)$$

And similarly, letting x_{i2} denote the set of household characteristics in round 2 that are observed in both the round 1 and round 2 surveys, the linear projection of round 2 consumption or income, y_{i2} onto x_{i2} is given by:

$$y_{i2} = \beta_2' x_{i2} + \varepsilon_{i2} \qquad (2)$$

Let z_1 and z_2 denote the poverty line in period 1 and period 2 respectively. Then to estimate the degree of mobility in and out of poverty we are interested in knowing, for example, what fraction of households in the population is above the poverty line in round 2 after being below the poverty line in round 1. That is, we are interested in estimating:

$$P(y_{i1} < z_1 \; and \; y_{i2} > z_2) \qquad (3)$$

which represents the degree of movement out of poverty for households over the two periods. However, the prime difficulty facing us with repeated cross-sections is that we do not know y_{i1} and y_{i2} for the same households. Without imposing a lot of structure on the data generating processes, one cannot point-identify the probability in (3). But it is possible to obtain bounds. To derive these bounds, note that we can rewrite this probability as:

$$P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1} \; and \; \varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \qquad (4)$$

We see that this probability depends on the joint distribution of the two error terms ε_{i1} and ε_{i2}, capturing the correlation of those parts of household consumption in the two periods which are unexplained by the household characteristics x_{i1} and x_{i2}. Intuitively, mobility will be greater the less correlated are ε_{i1} and ε_{i2}; household consumption in one period will be less

[4] Moreover, if surveys ask about when individuals developed chronic illnesses, or became unemployed, or suffered other such shocks which are correlated with poverty status, then these variables could also be included in *x*.

associated with that in the other period. One extreme case thus occurs when the two error terms are completely independent of each other. Another extreme case occurs when these two error terms are perfectly correlated.

To further operationalize the probability in (4), we make the following two assumptions.[5]

Assumption 1**:** *The underlying population sampled is the same in survey round 1 and survey round 2.*

In the absence of actual panel data on household consumption, this assumption ensures that we can use time-invariant household characteristics that are observed in both survey rounds to obtain predicted household consumption. Given that the underlying population being sampled in survey rounds 1 and 2 are the same, the time-invariant household characteristics in one survey round would be the same as in the other round, thus providing the crucial linkage between household consumption between the two periods. In other words, households in period 2 that have similar characteristics to those of households in period 1 would have achieved the same consumption levels in period 1 or vice versa.

Assumption 1 will not be satisfied if the underlying population changes through births, deaths, or migration out of sample, which could happen if the two survey periods are particularly far apart in time or as a result of major events, such as natural disasters or a sudden economic crisis, affecting the whole economy between the survey rounds. Assumption 1 may also not be satisfied due to survey-related technical issues such as changes in sampling methodology from one round to the next.[6]

Assumption 2: *The correlation* $\boldsymbol{\rho}$ *of* $\boldsymbol{\varepsilon_{i1}}$ *and* $\boldsymbol{\varepsilon_{i2}}$ *is non-negative.*

This assumption is to be expected in most applications using household survey data for at least three reasons. First, if the error term contains a household fixed effect, then households which have consumption higher than we would predict based on their *x* variables in round 1 will

[5] In addition to these two assumptions, we also use the (popular) standard assumptions that household consumption aggregates are consistently constructed and comparable over the two periods.

[6] In practice one can carry out a number of checks to test whether this assumption appears to hold with the cross-sectional data at hand by examining whether the observable time-invariant characteristics of a cohort change significantly from one survey round to the next. McKenzie (2001) provides an illustration of this approach for pseudo-panel analysis of Taiwanese households.

also have consumption higher than we would predict based on their *x* variables in round 2. Second, if shocks to consumption or income (for example, finding or losing a job) have some persistence, and consumption reacts to these income shocks, then consumption errors will also exhibit positive autocorrelation.

And finally, while for particular households we might see some negative correlation in incomes over time, the kind of factors leading to such a correlation are unlikely to apply to an entire population at the same time. For example, a household which lacks access to credit may cut expenditure in round 1 in order to pay for a wedding in round 2. For such a household we would see a lower consumption than their *x* variables would predict in round 1, and higher consumption than would be predicted for round 2. But this is unlikely to occur for the majority of households at the same time. Indeed, we will show this using panel data from several countries used in our analysis.

As in standard pseudo panel analysis these two assumptions will be best satisfied by restricting attention to households headed by people aged, say, 25 to 55. Analysis of mobility among households headed by those younger than 25 or older than 55 or 60 is more difficult since at those ages households are often beginning to form, or starting to dissolve. If income can be measured at the individual level, this may be less of a concern for individual income mobility than for household consumption mobility.

Given these two assumptions, we propose the following two theorems that provide the lower and upper bound estimates for poverty mobility. Since poverty immobility (i.e. households have the same poverty status in both survey rounds) is the opposite of poverty mobility, two closely related corollaries based on these two theorems provide the lower bound and upper bound of poverty immobility.

Theorem 1

The upper bound estimates of poverty mobility are given by the probability in expression (4) when the two error terms ε_{i1} and ε_{i2} are completely independent of each other, which implies $corr(\varepsilon_{i1}, \varepsilon_{i2}) = 0$. Specifically, the upper bound estimates of poverty mobility are given by

$$P\left(y_{i1}^{2U} < z_1 \; and \; y_{i2} > z_2\right) = P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \qquad (5)$$

for movements out of poverty, and

$$P\left(y_{i1}^{2U} > z_1 \; and \; y_{i2} < z_2\right) = P(\varepsilon_{i1} > z_1 - \beta_1' x_{i2}) P(\; \varepsilon_{i2} < z_2 - \beta_2' x_{i2}) \qquad (6)$$

for movements into poverty; where $y_{i1}^{2U} = \beta_1^{'} x_{i2} + \varepsilon_{i1}$ *and for* y_{i1}^{2U} *the superscript 2 stands for estimated round 1 consumption for households sampled in round 2, and U stands for the upper bound estimates of poverty mobility.*

Corollary 1.1

The biases for the upper bound estimates of poverty mobility in equations (5) and (6) above are respectively given by

$$Bias\; for\; P\left(y_{i1}^{2U} < z_1 \; and \; y_{i2} > z_2\right) = P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\; \varepsilon_{i2} > z_2 - \beta_2' x_{i2} | \varepsilon_{i1} \geq z_1 - \beta_1' x_{i2}) \quad (7)$$

$$Bias\; for\; P\left(y_{i1}^{2U} > z_1 \; and \; y_{i2} < z_2\right) = P(\varepsilon_{i1} > z_1 - \beta_1' x_{i2}) P(\; \varepsilon_{i2} < z_2 - \beta_2' x_{i2} | \varepsilon_{i1} \leq z_1 - \beta_1' x_{i2}) \quad (8)$$

Corollary 1.2

The lower bound estimates of poverty immobility are given by

$$P\left(y_{i1}^{2U} > z_1 \; and \; y_{i2} > z_2\right) = P(y_{i2} > z_2) - P\left(y_{i1}^{2U} < z_1 \; and \; y_{i2} > z_2\right) \qquad (9)$$

for households staying out of poverty in both rounds, and

$$P\left(y_{i1}^{2U} < z_1 \; and \; y_{i2} < z_2\right) = P(y_{i2} < z_2) - P\left(y_{i1}^{2U} > z_1 \; and \; y_{i2} < z_2\right) \qquad (10)$$

for households staying in poverty in both rounds.

Proof

See Appendix 1.

Theorem 2

The lower bound estimates of poverty mobility are given by the probability in expression (4) when the two error terms ε_{i1} *and* ε_{i2} *are identical (equal to each other), which implies* $corr(\varepsilon_{i1}, \varepsilon_{i2}) = 1$*. Specifically, the lower bound estimates of poverty mobility are given by*

$$P(y_{i1}^{2L} < z_1 \; and \; y_{i2} > z_2) = P(\varepsilon_{i2} < z_1 - \beta_1' x_{i2}) - P(\; \varepsilon_{i2} \leq z_2 - \beta_2' x_{i2}) \qquad (11)$$

for movements out of poverty, and

$$P(y_{i1}^{2L} > z_1 \; and \; y_{i2} < z_2) = P(\; \varepsilon_{i2} < z_2 - \beta_2' x_{i2}) - \; P(\varepsilon_{i2} \leq z_1 - \beta_1' x_{i2}) \qquad (12)$$

for movements into poverty; where $y_{i1}^{2L} = \beta_1^{'} x_{i2} + \varepsilon_{i2}$ *and for* y_{i1}^{2L} *the superscript 2 stands for estimated round 1 consumption for households sampled in round 2, and L stands for the lower bound estimates of poverty mobility .*

Corollary 2.1

The biases for the lower bound estimates of poverty mobility in equations (11) and (12) above are respectively given by

$$Bias\ for\ P(y_{i1}^{2L} < z_1\ and\ y_{i2} > z_2) = 1 - P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2} \cup \varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \qquad (13)$$

$$Bias\ for\ P(y_{i1}^{2L} > z_1\ and\ y_{i2} < z_2) = 1 - P(\varepsilon_{i1} > z_1 - \beta_1' x_{i2} \cup \varepsilon_{i2} < z_2 - \beta_2' x_{i2}) \qquad (14)$$

Corollary 2.2

The upper bound estimates of poverty immobility are given by

$$P(y_{i1}^{2L} > z_1\ and\ y_{i2} > z_2) = P(y_{i2} > z_2) - P(y_{i1}^{2L} < z_1\ and\ y_{i2} > z_2) \qquad (15)$$

for households staying out of poverty in both rounds, and

$$P(y_{i1}^{2L} < z_1\ and\ y_{i2} < z_2) = P(y_{i2} < z_2) - P(y_{i1}^{2L} > z_1\ and\ y_{i2} < z_2) \qquad (16)$$

for households staying in poverty in both rounds.

Proof
See Appendix 1.

The methods developed here aim to estimate the same level of movements into and out of poverty that one would observe in the genuine panel. Of course some of the mobility in the genuine panel data is spurious, arising from measurement error. There are several approaches in the existing literature for ways to correct mobility measures for such measurement error (e.g. Glewwe, 2010; Antman and McKenzie, 2007; Fields et al. 2007). The basic idea underlying all of these approaches is to study the mobility of some underlying variable—such as health, cohort characteristics, or assets—which is analogous to studying only the mobility which comes from the $\beta' x$ term and ignoring mobility which comes from ε.

While such an approach could be pursued here as well, it is not the purpose of our current exercise, which is to determine whether one can use repeated cross-sections to estimate the same level of mobility one sees in a panel, and whether the method is useful for showing which

characteristics are associated with more movements into and out of poverty. Note however that our estimates will still remain valid bounds for the true degree of mobility even under many types of measurement error, as stated in the theorem below.

Theorem 3

The lower bound and upper bound estimates of poverty mobility provided in Theorems 1 and 2 and Corollaries 1.2 and 2.2 are robust to classical measurement errors. The lower bound is also robust to general forms of non-classical measurement error, while the upper bound will still continue to be an upper bound in the presence of non-classical measurement error provided that this non-classical error does not cause assumption 2 to be violated.

Proof
See Appendix 1.

3. Non-parametric bounds

The theorems and corollaries in the previous section provide the theoretical framework for us to consider concrete procedures to estimate the lower and upper bounds of poverty mobility and immobility. This framework also shows that assumptions about the joint distribution for the two error terms are crucial for our estimates of poverty mobility, and there can be different approaches depending on different assumptions about this distribution. We consider two approaches to estimate the bounds on mobility: a non-parametric approach where we make no assumption about this joint distribution and then, in the next section, a parametric approach where we assume this joint distribution is bivariate normal. We start first with the non-parametric approach.[7]

3.1 Non-parametric Bounds

Upper-bound estimates for poverty mobility (and lower-bound estimates for poverty immobility)

We propose the following steps to obtain the quantities in (5), (6), (9) and (10)

[7] If we consider together the estimation method (OLS) and the distribution of the error term, perhaps it is more accurate to refer to this as a semi-parametric approach. However, we are using the terms "non-parametric" and "parametric" to highlight our assumptions about the distribution for the error terms. Also note that the phrases "upper bound" and "lower bound" pertain to their bounds on mobility, not to their bounds on levels of poverty.

Step 1: Using the data in survey round 1, estimate equation (1) and obtain the predicted coefficients $\hat{\beta}_1'$ and predicted residuals $\hat{\varepsilon}_{i1}$.

Step 2: For each household in round 2, take a random draw with replacement from the empirical distribution of the predicted residuals $\hat{\varepsilon}_{i1}$ obtained in step 1 and denote it by $\hat{\tilde{\varepsilon}}_{i1}$. Then using the data in survey round 2, the predicted coefficients $\hat{\beta}_1'$, and the residual $\hat{\tilde{\varepsilon}}_{i1}$, estimate, for each household in round 2, its consumption level in round 1, as follows

$$\hat{y}_{i1}^{2U} = \hat{\beta}_1' x_{i2} + \hat{\tilde{\varepsilon}}_{i1} \qquad (17)$$

Step 3: Estimate the quantities in (5), (6), (9) and (10), using $\hat{y}_{i1}^{2U}$ obtained from Step 2 above.

Step 4: Repeat steps 2 to 3 R times, and take the average of each quantity in (5), (6), (9) and (10) over the R replications to obtain the upper bound estimates of poverty mobility (or immobility). We use R= 500 in our simulations below.

Lower-bound estimates for poverty mobility (and upper-bound estimates for poverty immobility)

To obtain the lower bound estimates of the movement into and out of poverty for (3), we take the following steps

Step 1: Using the data in survey round 1, estimate equation (1) and obtain the predicted coefficients $\hat{\beta}_1'$. Then using the data in survey round 2, estimate equation (2) and obtain the residuals $\hat{\varepsilon}_{i2}$.

Step 2: Then using the data in survey round 2, the predicted coefficients $\hat{\beta}_1'$, and the residual $\hat{\varepsilon}_{i2}$, estimate the consumption level in round 1 for each household in round 2 as follows

$$\hat{y}_{i1}^{2L} = \hat{\beta}_1' x_{i2} + \hat{\varepsilon}_{i2} \qquad (18)$$

Step 3: Estimate the quantities in (11), (12), (15) and (16) using $\hat{y}_{i1}^{2L}$ obtained from Step 2 above.

A couple of remarks are in order about the above procedures. First, the bootstrapping of the error terms for the upper bound estimates is based on the condition of independence for the two error terms ε_{i1} and ε_{i2} as stated in Theorem 1. Second, unlike the upper bound estimates, the procedure for obtaining the lower bound estimates does not require repeating steps 2 to 3 R times since we are using each household's own predicted errors. And finally, we do not have to restrict estimation of predicted household consumption to the data in the second survey round (Steps 2 above) but can also use the data in the first survey round since the following identity always holds $P(y_{i1} < z_1 \text{ and } y_{i2} > z_2) \equiv P(y_{i2} > z_2 \text{ and } y_{i1} < z_1)$.[8]

3.2. Sharpening the Non-parametric Bounds

From Corollary 1.1, we see that the bias for our upper bound estimate of the probability a household is poor in the first period but non-poor in the second period is given by $P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2} | \varepsilon_{i1} \geq z_1 - \beta_1' x_{i2})$. Other things being equal, this probability will be smaller the greater is the variation in y_{i1} that can be explained by the set of variables in the vector *x*, and the lower the variation left to be represented by the error terms ε_{i1} and ε_{i2}. In particular, a weaker correlation between these error terms will tend to decrease the second term in this bias. Similarly, Corollary 2.1 also indicates that a weaker correlation between the error terms ε_{i1} and ε_{i2} will also tend to increase the second terms in (15) and (16) and thus decrease the overall biases.

This is equivalent to obtaining a high R^2 in the regression of y_{i1} on *x*. We can increase this R^2 and narrow the bounds by including a host of time-invariant (or deterministic) household characteristics. In addition, one can control for detailed geographic variables or region fixed effects. Taken together, a combination of household and regional characteristics may control for shocks which occur in particular regions or for people of particular characteristics, and may allow one to span household fixed effects. We shall see how well this strategy works in our empirical application in the next section.

3.3. Datasets

[8] If one wants to get standard errors for these bounds, then a bootstrap approach can be used. This would involve bootstrap resampling from the original cross-sections (taking account of survey weights) and then running the method described above within each bootstrap sample.

To examine how well our method performs in practice we implement our procedure using genuine panel data from Vietnam and Indonesia. Our two main data sets are the Vietnam Household Living Standards Surveys (VHLSSs) and the Indonesian Family Life Surveys (IFLSs). We use the VHLSSs in 2006 and 2008, which are nationally representative surveys implemented by Vietnam's General Statistical Office (GSO) with technical assistance from the World Bank. The VHLSSs are similar to the LSMS-type (Living Standards Measurement Survey) surveys supported by the World Bank in a number of developing countries and provide detailed information on the schooling, health, employment, migration, and housing, as well as household consumption and ownership of a variety of household durables for 9,189 households across the country in each round. These surveys are widely used in poverty assessment by the government and the donor community in Vietnam. One particular feature with these surveys is a rotating panel module, which collects panel data for one half of each survey round between two adjacent years. This combination of both cross-sectional data and panel data in one survey provides a perfect setting for us to validate our method.

Our data for Indonesia come from the Indonesian Family Life Surveys that were fielded by the RAND Corporation as part of their Labor and Population Program in collaboration with UCLA and the University of Indonesia. We use the IFLS2 and IFLS3 rounds corresponding to respectively, 1997 and 2000. The IFLS2 interviewed 7,500 households and the IFLS3 survey interviewed 10,400. The IFLS surveys are remarkable in the extent to which efforts were made to follow households over time. The IFLS2 and IFLS3 managed to resurvey 94.4 and 95.3%, respectively, of the original 7224 households interviewed in 1993 for the IFLS1 round. As is the case for the VHLSS, the IFLS surveys are multipurpose surveys that collect detailed information on a range of different topics – thereby permitting analysis of interrelated issues that single-purpose surveys do not. Information on economic outcomes like income and labor market outcomes can be combined with information on health outcomes, education and a whole host of additional socioeconomic indictors. Finally, in 1997, the IFLS fielded, alongside the IFLS2 household survey, a community survey about respondents' communities and public and private facilities. The analysis below draws on both household and community level information.

Since the IFLSs are panel surveys, we split the IFLS panels into two randomly drawn sub-samples (each representing half of the total sample), and we do the same for the VHLSS

panel component.[9] Call these sub-samples A and B respectively. Then we can use sub-sample A in the first round and sub-sample B in the second round as two repeated cross-sections which we then carry out our method on. We can then compare the mobility results obtained from using sub-sample A to impute round 1 values for sub-sample B to the results we would get using the genuine panel for sub-sample B. And we use panels with the same heads only for the genuine panels.

For our basic analysis we use the national poverty line in Vietnam provided with the VHLSSs (corresponding to D 2,559,850, and D 3,358,118 respectively for 2006 and 2008 (Glewwe, 2009)), and the Tornquist poverty line in the IFLS dataset (corresponding to Rp 86,128.1 in 2000 prices).[10] We show later in the paper that our results are robust to the choice of poverty line used.

3.4. Variable Choice

Our approach is built on a linear projection of consumption in round 1 onto individual, household and community-level characteristics that are also present in the data for round 2. As described in Elbers, Lanjouw and Leite (2009) in regard to poverty-mapping procedures, there is no obvious theory to guide the specification of what is essentially a forecasting model. However, certain diagnostics can be looked to for guidance. In general one would want to look well beyond explanatory power (a higher R^2 would tend to reduce the variance of the prediction error) to consider also statistical significance of the parameter estimates $\hat{\beta}_1$ (in order to reduce model error and the resultant overstatement of mobility) and to pay attention as well to concerns about over fitting. In the literature on poverty mapping, regressors have typically been drawn from several broad classes of variables including demographic variables (household size, gender and age profiles of households, etc.); human capital variables; labor market variables (occupational profiles), access to basic services and infrastructure (electricity access, connection to a piped water network, etc.); housing quality variables; ownership of durables; and community and locality-level variables.

[9] We only use the VHLSS panel component for non-parametric estimates to illustrate our method. For the parametric estimation in the next section, we construct our estimates using the VHLSS cross section component and then compare to the VHLSS panel component.

[10] We thank Kathleen Beegle and Kristin Himelein for help with the IFLS data.

Central to the present application of this approach is the additional requirement that regressors in these models be time invariant Obvious candidates are the ethnic, religious, or social-group membership of the household head. Other time-invariant variables can be readily constructed from the data, such as whether the household head was aged 15 or higher and educated at the primary school level by a particular moment in time. When retrospective data are collected, the range of time-invariant variables can be greatly expanded. For example, if both the 1997 and 1992 surveys collect information on whether the household had a fridge in 1992, this time-invariant variable can be used in the prediction models. Some retrospective variables, such as place of residence at the time of the last survey, are reasonably common in cross-sectional surveys, while other variables, such as sector of work, education level, and occupation at the time of the past survey, could easily be collected retrospectively. Context will also determine the choice of variables to use. If the main interest is on mobility in rural farming areas, one could presumably ask retrospective questions about land and major livestock holdings, and also condition on time-varying environmental variables like rainfall.

In our empirical applications below, we thus consider a hierarchy of six classes of prediction models which progressively employ more and more data that is sometimes, but not always, collected retrospectively. Since we have the actual panel data to work with, we can "force" regressors in round 2 to be time-invariant by using the round 1 values of selected variables. Clearly in a real-world application we would be dependent only on those variables collected during the second round, and would be concerned about possible recall error. But for the purpose of illustration here, we select variables we believe are likely to be recalled fairly accurately, and which could be asked retrospectively.[11]

The six models are built up progressively as follows:

1. (Basic Model) We begin with a sparse model, including only variables that can be readily judged as time-invariant. For example, we can include such regressors as the gender of the head, age of the household head (defined in round 1 year), birthplace of the head (rural/urban), whether the head ever attended primary school (or the head's completed

[11] In section 4 below, where we analyze the parametric variant of our approach, we wish to explore the scope for narrowing bounds via the imposition of additional structure and assumptions. In doing so we confine our attention to a basic model specification that can be readily estimated with currently available cross-section data.

years of schooling), the education level of the head's parents, and the head's religion and ethnicity.

2. We then introduce locational dummies such as urban/rural, or regional, dummies to measure where the household was living at the time of the first round survey. Most multipurpose surveys with a migration module would collect the information needed to allow these variables to be constructed, and even without a specific migration module, it is common to ask where households were living five years ago.[12]
3. Next, "community" variables are added, which can be obtained from community modules in most household surveys or perhaps population censuses. Once the retrospective location is identified (as per model 2), the use of such variables depends only on the availability of such auxiliary data, and not on further recall per se. In the case of Indonesia, these come from the community-level survey from 1997 and are inserted into both the IFLS2 and IFLS3 household surveys. For Vietnam, unfortunately the community module only collects data on rural communes, which can reduce the estimation sample size significantly. Thus we will use instead a household-level variable which indicates household poverty status as classified by the government in the first survey round.
4. We then add variables describing a household head's sector of work. At this point we clearly start to lean more heavily on our ability to explicitly insert round 1 values of these variables into the round 2 data. However, information on these variables could probably be easily collected on a retrospective basis. Indeed retrospective work histories have been collected in a number of labor surveys.
5. Further demographic variables that we force to be time-invariant are then added - such as household size and the number of children aged under 5. These would possibly be more difficult to collect retrospectively if household composition is very fluid, especially if the time interval between survey rounds increased. Nonetheless, it is not uncommon for surveys with a migration focus to ask about all individuals who have lived in the household in the past five years, and our impression is that households in many societies are able to recall such information relatively accurately.

[12] For example, Smith and Thomas (2003) find that Malaysian households can accurately recall migration histories, particularly for moves which are not very local or very short in duration.

6. (Full model) Finally, we include a number of variables describing a household's assets and housing quality at the time of round 1 - such as ownership of specific consumer durables like a TV and motorcycle, and the type of roofing and flooring material the household had. Including these variables increases the predictive power of the consumption models significantly. Such variables are not commonly collected in retrospective fashion in large multipurpose surveys, but they have been collected in some specific survey contexts.[13]

We estimate these models for log consumption per capita. We only use levels of the variables indicated above, but one could additionally enrich the models by including interactions (e.g. allowing the predictive impact of education for consumption to vary with region, sex of household head, etc.). The precise regression results used for the upper and lower bound estimates for model 1 (the "basic model") and model 6 (the "full model") for household consumption in the first period are presented in Tables 2.1a and 2.1b in Appendix 2.

3.5. Estimation Results

We turn, now, to one of the central questions in our study, namely whether analysis of duration of poverty, and mobility in and out of poverty, based on our synthetic panel data, can deliver results approximating the findings one would obtain with genuine panel data.[14] Table 1 presents our results. As we expected, the lower bound estimates underestimate mobility (understating movements into and out of poverty and overstating the extent to which people remain poor or remain non-poor) and the upper bound estimates overestimate mobility. The "truth" (true rate) tends to lie about midway between these bounds. We find thus that our approach does indeed present bounds within which the "truth" can be observed.[15]

[13] For example, de Mel, McKenzie and Woodruff (2009) ask Sri Lankan business owners and wage workers questions on whether their family owned a bicycle, radio, telephone, or vehicle when they were aged 12, and on the floor type their household had then. Individuals were able to recall such information relatively easily, although further work is needed to test how accurate such recall is. Berney and Blane (1997) offer some encouraging findings from a small sample in the U.K., showing high accuracy recall of toilet facilities, water facilities, and number of children in the household over a 50-year recall period.

[14] We refer to "synthetic panels" in our approach in an effort to distinguish our household-level analysis from the broader literature that works with cohort-means.

[15] Estimation is very similar when we obtain predicted household consumption on data from the first survey round instead of the second survey round. Thus for both the non-parametric and parametric estimates (in the next section), we only show results obtained on data from the second survey round.

What is particularly encouraging is that the width of these bounds is fairly reasonable. For example, using the full model, our bounds would suggest that between 3 and 10 percent of households in Indonesia, and between 3 and 7 percent of households in Vietnam moved out of poverty between the two rounds. Analysis based on the genuine panel data suggests that the true rates are well captured in these ranges, even after we adjust for one to two standard errors to these rates.

The results also illustrate the importance of being able to fit more detailed models to predict consumption, with generally narrower bounds for the models with richer specifications than the basic model—which is to be expected given our discussion in the previous Section. For example, the bounds for the proportion of the population falling into poverty in Vietnam between 2006 and 2008 are (0.5-8.6) using the basic model, (2.8-8.5) using model 2, (3.0-7.8) using model 3, (2.3-7.2) using model 5, and (2.1-6.8) using the full model. Corresponding to these narrower bounds is respectively a steady increase in R^2 of 0.33, 0.49, 0.55, 0.60, and 0.71 and a similar constant decrease in the correlation coefficient ρ (which is always positive and consistent with our Assumption 2).

In both countries it is the inclusion of locational variables to get to model 2, retrospective demographic variables to get to model 5, and especially the inclusion of the retrospective household asset variables to get to the full model that most increase the share of variation explained by the regressors and the greatest reduction in the size of the bounds. Efforts to collect retrospective data so as to be able to enrich the model specification thus do appear to be important.[16] The basic model has less predictive power, leading to wider intervals.

4. Sharpening the Bounds Further through a Parametric Method

The non-parametric method introduced and explored above has the advantage of requiring few assumptions to obtain bounds on the degree of mobility and producing fairly encouraging results. However, while the rich sets of regressors as used in the estimates in Table 1 may offer some directions on future survey designs (as well as a good illustration of what is feasible with

[16] This accords well with experience of applying the Elbers et al. (2003) method for small-area estimation purposes to poverty mapping. In those applications the methodology pursued most closely resembles the "upper bound", "full", approach here, and it is generally found that predicted poverty rates (calculated in the population census) closely track survey estimates at the broad-stratum level (see Demombynes et al. 2004).

our method), these may not currently be available for most countries. Without such a full set of variables, the bounds provided by the basic models may be too wide to be of use for practical purposes.

We thus move from this "ideal" setting to the rather more prosaic real-world one where only a subset of the above-considered regressors exists. We explore a parametric variant to our basic approach and impose some structure on the error terms in order to sharpen our bounds on mobility. We work with only with the basic model specification (i.e., Model 1) introduced above, including, in addition one dummy variable indicating urban or rural area of residence (and also show the non-parametric estimates for this specification).We now also estimate our models using only the cross-sectional components of the survey data, and compare our estimates of mobility against the "true" estimates calculated from the panel components.

This model thus puts modest demands on the data and would likely be applicable in most household surveys. We show that by introducing a distributional assumption on the error terms, and additional information on the likely plausible range of autocorrelation in these error terms, we can produce narrower bounds on mobility. We start with the following additional assumption.

Assumption 3*: $\boldsymbol{\varepsilon_{i1}}$ and $\boldsymbol{\varepsilon_{i2}}$ have a bivariate normal distribution with correlation coefficient ρ and standard deviations $\boldsymbol{\sigma_{\varepsilon_1}}$ and $\boldsymbol{\sigma_{\varepsilon_2}}$ respectively.*

Log-normality is a reasonable and often used approximation for the distribution of income or consumption, so this condition may hold approximately in practice and can be checked, as will be illustrated in our empirical section.

4.1. Parametric Estimation Framework

Given Assumptions 1 and 3, it is straightforward to see that the percentage of households that are poor in the first period but nonpoor in the second period $P(y_{i1} < z_1 \text{ and } y_{i2} > z_2)$ can be estimated by

$$P^E(y_{i1} < z_1 \text{ and } y_{i2} > z_2) = P(\beta_1' x_{i2} + \varepsilon_{i1} < z_1 \text{ and } \beta_2' x_{i2} + \varepsilon_{i2} > z_2)$$
$$= \Phi_2\left(\frac{z_1 - \beta_1' x_{i2}}{\sigma_{\varepsilon_1}}, -\frac{z_2 - \beta_2' x_{i2}}{\sigma_{\varepsilon_2}}, -\rho\right) \quad (19)$$

where $\Phi_2(.)$ stands for the bivariate normal cumulative distribution function (cdf)) (and $\phi_2(.)$ stands for the bivariate normal probability density function (pdf)).

Since we know that for any x, y, and ρ, $\frac{\partial \Phi_2(x, y, \rho)}{\partial \rho} = \phi_2(x, y, \rho) > 0$ (Sungur, 1990), equation (19) indicates that the key difference between a household's true consumption level and its lower bound and upper estimates of mobility lies with the correlation term ρ. Since ρ is bounded by the interval [0, 1] (Assumption 2), and the correlation term in equation (19) above has a negative sign ($-\rho$), a lower value of ρ means a higher probability of entering/ exiting poverty (i.e., a higher degree of mobility or lower degree of immobility) in the second period and vice versa.

In fact, the non-parametric lower bound and upper bound estimates of poverty mobility correspond to assuming ρ being equal to its maximum value (1) and minimum value (0) respectively.[17] However, as was noted in our discussion of Table 1, the true value of ρ in all likelihood lies somewhere in between these two values of 0 and 1. If we can have a better estimate of ρ, we can narrow the gap between these lower bound and upper bound estimates of poverty mobility. Thus we can tighten Assumption 2 as follows.

Assumption **2'**: $\boldsymbol{\rho \in [\rho_S, \rho_H]}$ *where* $\boldsymbol{\rho_S}$ *is the smallest hypothesized value of* $\boldsymbol{\rho}$ *and* $\boldsymbol{\rho_H}$ *the highest hypothesized value, with* $\mathbf{0 < \rho_S < \rho_H < 1}$.

In searching for the range of appropriate values for ρ, there seem to be two options available: i) we can look at actual panel data in previous time periods from the same country (or for sub-samples of the data) or, ii) we can consider actual panel data in (say, economically or geographically) similar settings elsewhere. We will pursue this second option below and calculate a range of different values for ρ from a similar model specification estimated in a number of different countries for which panel data exist.

4.2. Parametric Estimation Procedures

[17] In particular, when $\rho = 0$ or $\rho = 1$, the parametric analogues of the upper and lower bound estimates of poverty mobility in (5), (6), (11) and (12) are obtained by replacing the general probability notation "P(.)" with the normal cdf $\Phi(.)$.

Upper-bound estimates for poverty mobility (and lower-bound estimates for poverty immobility)

We propose the following steps to obtain the quantities in (5), (6), (9) and (10)

Step 1: Using the data in survey round 1, estimate equation (1) and obtain the predicted coefficients $\hat{\beta}_1'$, and the predicted standard error $\hat{\sigma}_{\varepsilon_1}$ for the error term ε_{i1}. Using the data in survey round 2, estimate equation (2) and obtain similar parameters $\hat{\beta}_2'$ and $\hat{\sigma}_{\varepsilon_2}$.

Step 2: For each household in round 2, calculate the quantities in (5), (6), (9) and (10) as follows using the smallest hypothesized value of ρ, ρ_S

$$\hat{P}^{2U}(y_{i1} < z_1 \text{ and } y_{i2} < z_2) = \Phi_2\left(\frac{z_1 - \hat{\beta}_1' x_{i2}}{\hat{\sigma}_{\varepsilon_1}}, \frac{z_2 - \hat{\beta}_2' x_{i2}}{\hat{\sigma}_{\varepsilon_2}}, \rho_S\right) \quad (20)$$

$$\hat{P}^{2U}(y_{i1} < z_1 \text{ and } y_{i2} > z_2) = \Phi_2\left(\frac{z_1 - \hat{\beta}_1' x_{i2}}{\hat{\sigma}_{\varepsilon_1}}, -\frac{z_2 - \hat{\beta}_2' x_{i2}}{\hat{\sigma}_{\varepsilon_2}}, -\rho_S\right) \quad (21)$$

$$\hat{P}^{2U}(y_{i1} > z_1 \text{ and } y_{i2} < z_2) = \Phi_2\left(-\frac{z_1 - \hat{\beta}_1' x_{i2}}{\hat{\sigma}_{\varepsilon_1}}, \frac{z_2 - \hat{\beta}_2' x_{i2}}{\hat{\sigma}_{\varepsilon_2}}, -\rho_S\right) \quad (22)$$

$$\hat{P}^{2U}(y_{i1} > z_1 \text{ and } y_{i2} > z_2) = \Phi_2\left(-\frac{z_1 - \hat{\beta}_1' x_{i2}}{\hat{\sigma}_{\varepsilon_1}}, -\frac{z_2 - \hat{\beta}_2' x_{i2}}{\hat{\sigma}_{\varepsilon_2}}, \rho_S\right) \quad (23)$$

Lower-bound estimates for poverty mobility (and upper-bound estimates for poverty immobility)

Lower-bound estimates of poverty mobility (and upper-bound estimates for poverty immobility) can likewise be obtained by using the same steps with ρ_H in place of ρ_S.

Note that in the special case that the true value of ρ is somehow known, the bounds collapse to a point estimate. It is not unreasonable to think of possible scenarios where—say, to save costs—small but representative panel surveys were fielded and ρ estimated from such surveys could be combined with cross sectional surveys to estimate poverty transitions in the larger datasets.

As with the non-parametric case, it should be noted that we obtain the predicted parameters from both survey rounds and then calculate the poverty dynamics on data from the second survey round (x_{i2}), but we can also first obtain the predicted parameters from both survey rounds and then calculate the poverty dynamics on data from the first survey round (x_{i1}). The two approaches should give us the same results,[18] since the same identity holds as for the non-parametric estimation.

4.3. Parametric Estimation Results

Normality Assumptions and determining ρ

Since the key assumption required for our parametric approach is normality of the error terms in the regressions of household consumption on household (time-invariant) characteristics, we start off by plotting for each country and year the distribution for the estimated error terms (ε_i) against the normal distribution. A casual visual inspection indicates that the former (dotted line) closely resembles the latter (solid line) in each year (Appendix 2, Figure 2.1), although the graphs for Vietnam look somewhat better than those for Indonesia. However, formal multivariate normality tests (Doornik and Hansen, 2008) reject the assumption of normality distribution (univariate or bivariate) for the error terms in both countries. Despite this rejection we will maintain the assumption below, and thereby illustrate the performance of our parametric bounding methods in a typical practical situation where the underlying distributional assumption may not hold precisely.

[18] However, this variant approach results in changes to the bivariate probability formulas to calculate the poverty dynamics probabilities in equations (20)- (23), which are given below

$$\hat{P}^{2U}(y_{i2} < z_2 \text{ and } y_{i1} < z_1) = \Phi_2\left(\frac{z_1 - \hat{\beta}_1' x_{i1}}{\hat{\sigma}_{\varepsilon_1}}, \frac{z_2 - \hat{\beta}_2' x_{i1}}{\hat{\sigma}_{\varepsilon_2}}, \rho\right) \quad (20')$$

$$\hat{P}^{2U}(y_{i2} > z_2 \text{ and } y_{i1} < z_1) = \Phi_2\left(\frac{z_1 - \hat{\beta}_1' x_{i1}}{\hat{\sigma}_{\varepsilon_1}}, -\frac{z_2 - \hat{\beta}_2' x_{i1}}{\hat{\sigma}_{\varepsilon_2}}, -\rho\right) \quad (21')$$

$$\hat{P}^{2U}(y_{i2} < z_2 \text{ and } y_{i1} > z_1) = \Phi_2\left(-\frac{z_1 - \hat{\beta}_1' x_{i1}}{\hat{\sigma}_{\varepsilon_1}}, \frac{z_2 - \hat{\beta}_2' x_{i1}}{\hat{\sigma}_{\varepsilon_2}}, -\rho\right) \quad (22')$$

$$\hat{P}^{2U}(y_{i2} > z_2 \text{ and } y_{i1} > z_1) = \Phi_2\left(-\frac{z_1 - \hat{\beta}_1' x_{i1}}{\hat{\sigma}_{\varepsilon_1}}, -\frac{z_2 - \hat{\beta}_2' x_{i1}}{\hat{\sigma}_{\varepsilon_2}}, \rho\right) \quad (23')$$

where ρ is set to equal ρ_S and ρ_H respectively for the upper bound and lower bound estimates for poverty mobility.

We calculate different values for ρ using true panel data from several developing countries: Bosnia- Herzegovina, Indonesia, Lao PDR, Nepal, Peru, and Vietnam. Our estimates are provided in Table 2.[19]. Clearly, this list is far from being exhaustive—and we expect future research will build on this—but this sample of countries spans different regions and income levels at different points in time over the past decade. For these estimates, we use model specifications which are as similar as permissible by the data available to the basic model employed above for the non-parametric estimates plus a dummy variable indicating area of residence (urban/ rural). These are also the same model specifications we use for predictions using the cross sectional data.

The estimates in Table 2 show that ρ ranges from 0.39 (for Nepal during 1995-2004) to 0.66 (for Vietnam during 2004-2006) which is arguably a rather tight range compared to its theoretical range of [0, 1]. [20] However, to be on the safe side, we will widen this range a bit more and use the two pairs of values of (0.2, 0.8) and (0.3, 0.7) for our subsequent bound estimates.

Lower and Upper Bound Estimates

The lower bounds and upper bounds of poverty mobility for Vietnam and Indonesia are further examined in Table 3. Our bound estimates are considered in three model specifications: Specification 1 provides the most conservative bounds where ρ are respectively set to 1 and 0, and Specifications 2 and 3 provide less conservative bounds where ρ are respectively assumed to be equal to [0.8, 0.2] and [0.7, 0.3]. Clearly, the estimates from Specification 1 would be the parametric equivalence of our previous non-parametric estimates—which are also shown for comparison under the column "Non-parametric bound"—but we will focus here on the parametric estimates for interpretation. The bound estimates are expected to be sequentially tighter for Specifications 1, 2 and 3; however, this naturally comes with a trade-off since the tighter the bounds, the higher the chance that these bounds do not encompass the true rates.

[19] The data are from Bosnia- Herzegovina during 2001-2004 (Demirguc-Kunt, Klapper and Panos, 2009), Lao PDR during 2002-2007 (Lao Department of Statistics, 2009), Nepal during 1995-2004 (Nepal's Central Bureau of Statistics, 2004), and Peru during 2004-2006 (Peruvian Statistics Bureau—INEI). These countries' household surveys are similar to the LSMSs and thus can provide a relevant and comparable range of values for this correlation coefficient. In addition we also employ the 2004 VLHSS.

[20] These positive values for ρ confirm again the validity of our Assumptions 2 and 2'.

Table 3 shows that the true poverty dynamic rates obtained from the panel data are well within the lower and upper bounds respectively provided by Specification 1, which are very similar to those obtained by the non-parametric method. Notably, except for those remaining non-poor in both periods, these true poverty rates are also bounded by the less conservative estimates from Specification 2, which shrink the intervals between the lower and upper bound in Specification 1 by around half for both countries. For example, the proportion of households who were poor in 2006 but nonpoor in 2008 for Vietnam is 5.7 percent, which lies between the less conservative lower and upper bound estimates of [4.3, 8.5] under Specification 2. This interval width of 4.2 percent is half that of the most conservative bounds under Specification 1, which has interval [0.4, 9.4].

As expected, estimates under Specification 3 provide even a tighter range, but these bounds now do not contain the true rates not only for those remaining nonpoor in both periods, but also those falling into poverty in the second period for Vietnam and those remaining poor in both periods for Indonesia. The silver lining, however, is that the differences between the imprecise bounds and the true rates range from 0.3 to 0.9 percentage points (which are roughly 5 to 20 percent in relative terms), except for the estimates for those who remained non-poor in both periods. Even in these worst cases, the order of magnitude for the miscalculation only amounts to around 1 percent of the true rate for Vietnam (e.g., (82.3- 81.1)/ 82.3= 0.014) and 4 percent of the true rate for Indonesia. Moreover, the width of the intervals obtained is now typically less than one third of the corresponding intervals offered by Specification 1.[21]

5. Alternative Poverty Lines and Mobility Profiles

We examine in this section robustness to the choice of poverty line, and an extension of our analysis to subpopulation groups.

5.1. Robustness to Choice of Poverty Line

The preceding analysis has all been based on one particular poverty line. The question then arises as to whether the approach described here is also successful in bounding true mobility

[21] The estimates in Table 3 are obtained by applying the predicted coefficients and error terms from both survey rounds to data in the second survey round. Results are similar when we replicate these results using data in the first survey round. Results available on request.

when alternative poverty lines are considered. From the proofs offered in Appendix 1, there is no particular reason this should not be the case. However, as an empirical robustness check on the estimation, we consider different poverty lines. A related question is whether the tightness with which our bounds "sandwich" the truth is constant for different values of the poverty line. We investigate these questions by calculating upper and lower bounds on mobility, as well as the truth, for the set of poverty lines spanning the range of possible base year poverty rates from 0 to 100 percent using the non-parametric method. Figure 2 illustrate our results in terms of the fraction of the population who escape poverty for Indonesia.[22]

The IFLS "true" panel data indicate that the share of the population able to escape poverty is low when the base year poverty line (and hence aggregate poverty) are sufficiently low (Figure 1). As the poverty line increases in value, a larger share of the base year population is considered poor and the percent that escapes poverty also rises. As the poverty line continues to rise an increasing fraction of the base year population is counted as poor and eventually the share of that underlying population that manages to escape poverty starts to decline. When the line is sufficiently high the whole population is poor and remains poor. Figure 1 shows that the inverted U-curve pattern traced out by the IFLS panel data is tracked fairly closely by our lower and upper bound synthetic panel estimates of mobility out of poverty. Allowing for some overlap and crossing attributable to statistical uncertainty, the bounds do "sandwich" the truth over the full range of possible poverty lines. Figure 1 also illustrates that the gap between the upper and lower bound estimates is at its widest when around half of the base-year population is considered poor, and also the largest share of the population is able to escape poverty. At more extreme poverty lines, the bounds are much closer together, pointing also to much lower rates of mobility out of poverty.

Other figures considering poverty immobility (not shown) also provide similar results. In sum, our approach is found to work well for the full possible range of poverty lines that might be specified, and we find that our bounds are, indeed, upper and lower bounds to the "truth" irrespective of where the poverty line is drawn.

5.2. Poverty Transitions Among Population Sub-Groups

[22] Similar results for Vietnam are available upon request.

While our proposed bounds appear to work well for the whole population, it is of interest to investigate whether the same is true for smaller population groups for several reasons. First, in designing effective social safety nets, policy makers often focus on smaller but more disadvantaged groups, rather than the whole population. This is especially the case in developing countries where due to resources constraints, allocations must be prioritized. Second, due to cost and logistical considerations sample sizes of true panel data are often fairly small, and this limits their applicability to the assessment of mobility across small population groups. In cases where the sample sizes of panel data are too small, these data may offer either imprecise or even unreliable estimates due to large standard errors or the non-representativeness of the data themselves. One of the advantages of the approach considered here is that our synthetic panels are based on cross-sectional data which often comprise far larger samples; if the samples of our synthetic panels are large enough, estimates based on these synthetic panels may better represent the target population.[23]

We estimate and plot the proposed parametric bounds (using Specifications 1 and 2, Table 3) against the true poverty dynamic rates for sub-groups of the population in Vietnam categorized by ethnicity (i.e., ethnic minority groups), female-headed households, education achievement (i.e., primary education or higher, lower secondary education or higher), and residence areas (i.e., urban households or regions the household live in) in Figures 2 to 5. Clearly, these categorizations can overlap but they can provide a first cut at profiling poverty mobility for different groups. Except for a few cases (e.g., households living in the North Central in Figure 2 and Figure 3, in the Mekong Delta, North Central or Southeast regions in Figure 5), the true rates lie within the less conservative bounds. Again, for these exceptional cases where the bounds are off, the differences do not appear to be large either.

These graphs also indicate that ethnic minority groups are the group most vulnerable to chronic poverty (Figure 2) and have very high mobility both into and out of poverty (Figures 3 and 4).[24] The Northwest group has similar patterns with ethnic minority groups since the

[23] It is a well-known fact that while panel data may be representative of the whole population, they may not be representative of all sub-population groups. For an (extreme) example, most panel data can perhaps provide good estimates of income dynamics for the population that is literate, but may not be able to provide reliable estimates for the population that has a Ph.D. degree.

[24] See Dang (forthcoming) for a more detailed discussion of the welfare for ethnic groups in recent years in Vietnam.

majority of the population in this region (76%) belong to ethnic minority groups.[25] On the other hand, households living in the urban area or households with their heads having a lower secondary education or higher appear to be better off than most other groups in the country.

Again, these evaluations of our bounds are only predicated on the assumptions that these small but true panel data are representative of the target population; otherwise, we may simply use estimates from the synthetic panels because of their larger sample sizes and supposedly better representativeness.

6. Conclusions and Future Directions

Genuine panel data are still rare in the developing world, and when they are available, the samples are often relatively small, with limited or infrequent duration, and in some cases, occur with significant attrition. This has limited the feasibility of constructing even the most simple descriptions of movements in and out of poverty for most countries. Yet policymakers and researchers do care about such movements, and most countries do field repeated cross-sectional surveys of income or consumption on a reasonably regular basis. In this paper we have developed a method for using existing cross-section data to provide some bounds on the extent of movements into and out of poverty, and results from both Indonesia and Vietnam suggest these bounds can be made narrow enough in practice to make the estimates useful.[26]

The success of the method depends on either how well one can predict the dependent variable of interest (for the non-parametric approach) or how well we can capture the range of autocorrelation for the error terms (for the parametric approach). For the former in the case of consumption or income dynamics, we have found that our accuracy in doing this, and the resulting width of the bounds for mobility, is significantly better when we are able to use retrospective information on the demographic composition of the household, the ownership of consumer durables and basic housing materials. Such variables are typically collected only concurrently, and not retrospectively, in most household surveys. It could also be promising to ask questions on when certain shocks such as development of chronic illness or death of a spouse

[25] Authors' calculation from the 2008 VHLSS.

[26] Preliminary evidence to support this can be seen by new efforts underway to use the methodology developed in this paper to systematically examine poverty dynamics in a number of Latin American countries. This work is being carried out by the World Bank's Latin American and the Caribbean office, not the authors of this study.

occur, since such variables might also help predict poverty status. Since it is certainly much less costly to collect this information than it is to field panel surveys, our results suggest it might be worth experimenting with the inclusion of such questions in some upcoming nationally representative surveys in order to be able to provide basic estimates of poverty transitions.

While better predicted household consumption would clearly improve parametric estimates as well, for the latter, we note that the empirically relevant ranges for the correlation term ρ would likely vary for different welfare outcomes (those for, say, household consumption can clearly differ from those for employment). Future research could thus focus on extending the list of empirically estimated correlation terms by looking at panel data from different countries, as well as creating a similar list for other welfare outcomes. These typologies of the range of autocorrelation for the error terms could then be used to provide estimates for countries with similar settings. Another promising direction is to collect data on a smaller subpanel (i.e., for cost savings) and combine the estimated correlation terms from this subpanel with the larger sample-sized cross sections to estimate poverty mobility.

References

Antman, Francisca and David McKenzie (2007) "Earnings Mobility and Measurement Error: A Synthetic panel Approach", *Economic Development and Cultural Change* 56(1): 125-162.

Banks, James, Richard Blundell, and Agar Brugiavini. (2001). "Risk Pooling, Precautionary Saving and Consumption Growth". *Review of Economic Studies*, 68(4): 757-779.

Berney, L.R. and D.B. Blane (1997) "Collecting Retrospective Data: Accuracy of recall after 50 years judged against historical records", *Social Science and Medicine* 45(10): 1519-25.

Casella, George and Roger L. Berger. (2002). *Statistical Inference*, 2nd Edition. California: Duxbury Press.

Dang, Hai-Anh. (forthcoming). "Vietnam: A Widening Poverty Gap for Ethnic Minorities", in Gillette Hall and Harry Patrinos. (Eds.) "*Indigenous Peoples, Poverty and Development*". Cambridge University Press.

Deaton, Angus (1985) "Panel Data from Time Series of Cross-Sections", *Journal of Econometrics* 30: 109-216.

Deaton, Angus and Christina Paxson. (1994). "Intertemporal Choice and Inequality". *Journal of Political Economy*, 102(3): 437- 467.

De Mel, Suresh, David McKenzie, and Christopher Woodruff (2010) "Who are the microenterprise owners? Evidence from Sri Lanka on Tokman v. de Soto", pp.63-87 in Joshua Lerner and Antoinette Schoar (eds.) *International Differences in Entrepreneurship*. NBER, Cambridge, MA.

Demirguc-Kunt, Asli, Leora F. Klapper, and Georgios A. Panos. (2009). "Entrepreneurship in Post-Conflict Transition: The Role of Informality and Access to Finance". Policy Research Working Paper 4935, DECRG, The World Bank.

Demombynes, G., Elbers, C., Lanjouw, J., Lanjouw, P., Mistiaen, J. and Ozler, B. (2004) 'Producing a Better Geographic Profile of Poverty: Methodology and Evidence from Three Developing Countries'. In Shorrocks, A. and van der Hoeven, R. (eds) *Growth, Inequality and Poverty* (Oxford University Press).

Elbers, C., Lanjouw, J.O, and Lanjouw, P. (2002) "Micro-Level Estimation of Welfare" Policy Research Working Paper 2911, DECRG, The World Bank.

Elbers, C. Lanjouw, J.O. and Lanjouw, P. (2003) "Micro-level Estimation of Poverty and Inequality" *Econometrica*, 71(1): 355-364.Elbers, C. Lanjouw, P. and Leite, P. (2010) 'Brazil Within Brazil: Testing the Poverty Map Methodology in Minas Gerais', mimeo, DECRG, the World Bank.

Fields, Gary, Robert Duval-Hernández, Samuel Freije Rodríguez, and María Laura Sánchez Puerta. (2007). "Earnings Mobility in Argentina, Mexico, and Venezuela: Testing the Divergence of Earnings and the Symmetry of Mobility Hypotheses." Mimeo. School of Industrial and Labor Relations, Cornell University.

Foster, James E. (2009) "A Class of Chronic Poverty Measures", pp.59-76 in Tony Addison, David Hulme, and Ravi Kanbur. (eds.) *Poverty Dynamics: Interdisciplinary Perspectives*. Oxford University Press: New York.

Gibson, John (2001) "Measuring Chronic Poverty Without a Panel", *Journal of Development Economics* 65(2): 243-66.

Glewwe, Paul (2009). "*Mission Report for Trip to Vietnam June 5-16, 2009*". Reported submitted to the World Bank.

___. (2010). "How Much of Observed Mobility is Measurement Error? IV Methods to Reduce Measurement Error Bias, with an Application to Vietnam", Mimeo. University of Minnesota.

Güell, Maia and Luojia Hu. (2006). "Estimating the Probability of Leaving Unemployment Using Uncompleted Spells from Repeated Cross-Section Data". *Journal of Econometrics*, 133: 307–341.

Hojman, Daniel and Felipe Kast. (2009). "On the Measurement of Poverty Dynamics", Working Paper Series RWP09-035, John F. Kennedy School of Government, Harvard University.

McKenzie, David (2001) "Consumption Growth in a Booming Economy: Taiwan 1976-96", *Yale University Economic Growth Center Discussion Paper no. 823.*

Pencavel, John. (2007). "A Life Cycle Perspective on Changes in Earnings Inequality among Married Men and Women". *Review of Economics and Statistics*, 88(2): 232-242.

Smith, James P. and Duncan Thomas (2003) "Remembrance of Things Past: Test-retest reliability of retrospective migration histories", *Journal of the Royal Statistical Society Series A*, 166(1): 23-49.

Sungur, Engin A. (1990). "Dependence Information in Parameterized Copulas". *Communications in Statistics- Simulation and Computation*, 19: 4, 1339 — 1360.

Verbeek, Marno (2008) "Synthetic panels and repeated cross-sections", pp.369-383 in L. Matyas and P. Sevestre (eds.) *The Econometrics of Panel Data*. Springer-Verlag: Berlin.

Table 1: Poverty Dynamics from Synthetic Panel Data and Actual Panel Data for Indonesia and Vietnam

Country	Poverty status	Non-parametric lower bound						Truth	Non-parametric upper bound					
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6		Model 6	Model 5	Model 4	Model 3	Model 2	Model 1
Indonesia 1997-2000	Poor, Poor	12.8	12.1	11.9	11.1	11.8	11.7	5.9	4.2	3.6	3.0	3.0	2.9	2.9
								(0.4)						
	Poor, Nonpoor	1.2	1.4	1.4	2.0	2.6	3.2	8.1	10.3	10.2	10.8	10.9	10.8	11.1
								(0.5)						
	Nonpoor, Poor	1.7	2.4	2.5	3.4	2.7	2.8	7.9	10.3	10.9	11.5	11.5	11.6	11.6
								(0.5)						
	Nonpoor, Nonpoor	84.3	84.1	84.1	83.5	82.9	82.3	78.1	75.2	75.3	74.8	74.6	74.7	74.4
								(0.7)						
	ρ	0.54	0.529	0.521	0.521	0.475	0.421							
	Adjusted R2	0.193	0.21	0.215	0.231	0.329	0.421							
	N	1638	1638	1638	1638	1638	1638	3517	1638	1638	1638	1638	1638	1638
Vietnam 2006-2008	Poor, Poor	12.5	10.2	10.1	10.1	10.8	11	7.6	6.3	5.9	5.2	5.2	4.6	4.5
								(0.5)						
	Poor, Nonpoor	0.4	2.6	2.6	2.7	3.3	3.3	5.7	6.6	7.3	7.3	7.4	8.5	9.4
								(0.4)						
	Nonpoor, Poor	0.5	2.8	3.0	3.0	2.3	2.1	4.4	6.8	7.2	7.9	7.8	8.5	8.6
								(0.4)						
	Nonpoor, Nonpoor	86.5	84.3	84.3	84.2	83.6	83.6	82.3	80.3	79.6	79.6	79.5	78.4	77.6
								(0.7)						
	ρ	0.654	0.584	0.554	0.547	0.516	0.394							
	Adjusted R2	0.334	0.494	0.548	0.559	0.60	0.71							
	N	1335	1335	1335	1335	1335	1335	2728	1335	1335	1335	1335	1335	1335
Note:	1.Poverty rates in percent are calculated using halves from the IFLS panel and the VHLSS panel component, and predictions obtained using data in the second survey rounds.													
	Full regression results are provided in Tables 2.1a and 2.1b in Appendix 2.													
	2. All numbers are weighted using population weights for each survey round. Standard errors in parentheses.													
	3. Number of replications for the estimates is 500.													
	4. Household heads' ages are restricted to between 25 and 55 in the first survey round.													

Table 2: Estimated ρ from Actual Panel Data for Different Countries

Country	Survey Year	ρ
Bosnia- Herzegovina	2001	0.43
	2004	
Indonesia	1997	0.47
	2000	
Lao PDR	2002-03	0.40
	2007-08	
Nepal	1995-96	0.39
	2003-04	
Peru	2004	0.58
	2006	
Vietnam	2004	0.66
	2006	
	2004	0.35
	2008	
	2006	0.62
	2008	

Note: 1. Each cell represents results from one regression, except for the cells under " ρ".

2. Household heads' ages are restricted to between 25 and 55 in the first survey round

3. ρ is the correlation coefficient between the error terms for the panel data.

Table 3: Poverty Dynamics from Synthetic Panel Data and Actual Panel Data for Indonesia and Vietnam

Country	Poverty status	Non-parametric bound	Parametric lower bound			Truth		Parametric upper bound		Non-parametric bound
			Spec. 1	Spec. 2	Spec. 3		Spec. 3	Spec. 2	Spec. 1	
Indonesia 1997-2000	Poor, Poor	13.3	15.9	11.1	9.8	5.9	6.1	5.4	4.0	3.3
						(0.4)				
	Poor, Nonpoor	1.6	1.7	6.5	7.8	8.1	11.5	12.2	13.5	12.3
						(0.5)				
	Nonpoor, Poor	0.9	0.9	5.7	7.0	7.9	10.7	11.5	12.8	11.7
						(0.5)				
	Nonpoor, Nonpoor	84.3	81.5	76.7	75.4	78.1	71.7	71.0	69.6	72.7
						(0.7)				
	N	1710	1710	1710	1710	3517	1710	1710	1710	1710
Vietnam 2006-2008	Poor, Poor	11.8	13.1	9.2	8.3	7.6	5.6	5.1	4.1	3.9
						(0.5)				
	Poor, Nonpoor	0.6	0.4	4.3	5.3	5.7	8.0	8.5	9.4	9.2
						(0.4)				
	Nonpoor, Poor	0.4	0.5	4.4	5.3	4.4	8.0	8.5	9.5	8.4
						(0.4)				
	Nonpoor, Nonpoor	87.2	86.0	82.1	81.1	82.3	78.4	77.9	77.0	78.6
						(0.7)				
	N	3701	3701	3701	3701	2728	3701	3701	3701	3701

Note: 1.Poverty rates in percent are calculated using halves from the IFLS panel and the VHLSS cross section component, and predictions obtained using data in the second survey rounds.

2. All numbers are weighted using population weights for each survey round. Standard errors in parentheses.

3. Specification 1 assumes $\rho = 1$ and $\rho = 0$ for the lower bounds and upper bounds respectively and is the parametric equivalence of the nonparametric bounds. Specification 2 approximates ρ with 0.8 and 0.2, and Specification 3 approximates ρ with 0.7 an 0.3 for the lower bounds and upper bounds respectively. Number of replications for non-parametric estimates is 500.

4. Household heads' ages are restricted to between 25 and 55 for the first survey round and between 27 and 57 for the second survey round.

Figure 1: Estimates of Mobility Out of Poverty for Alternative Poverty Lines, Indonesia

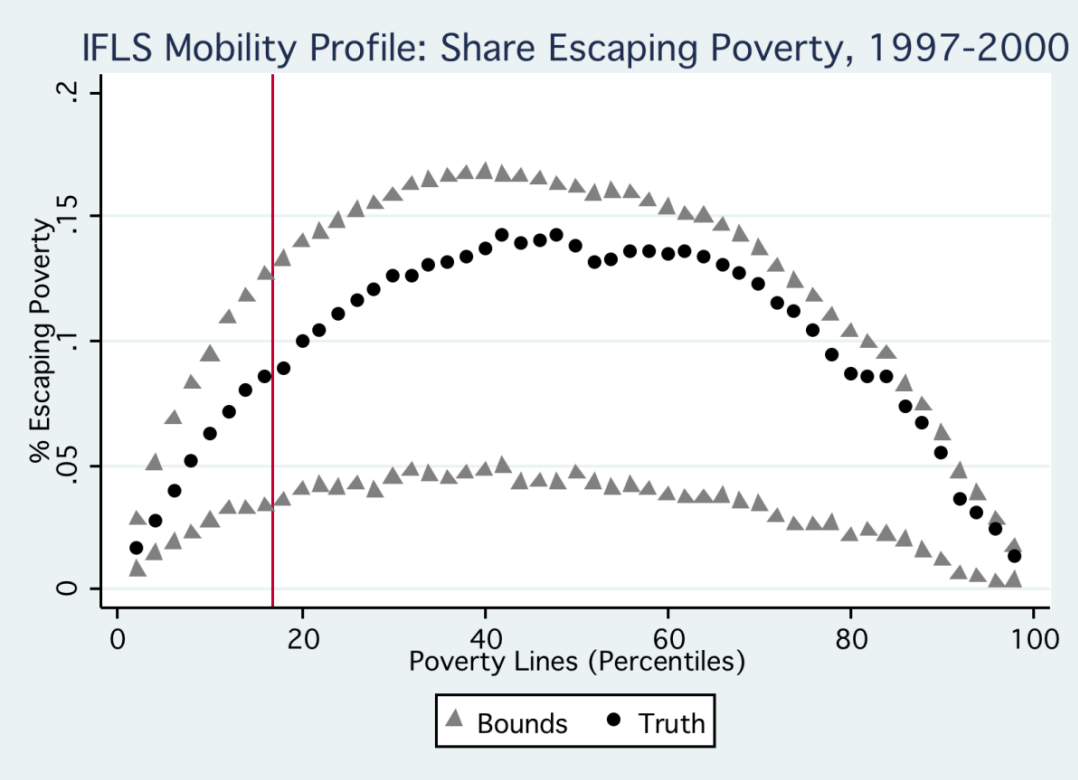


Figure 2: Profiles for Those Who Remained Poor in Both Periods, Vietnam 2006- 2008

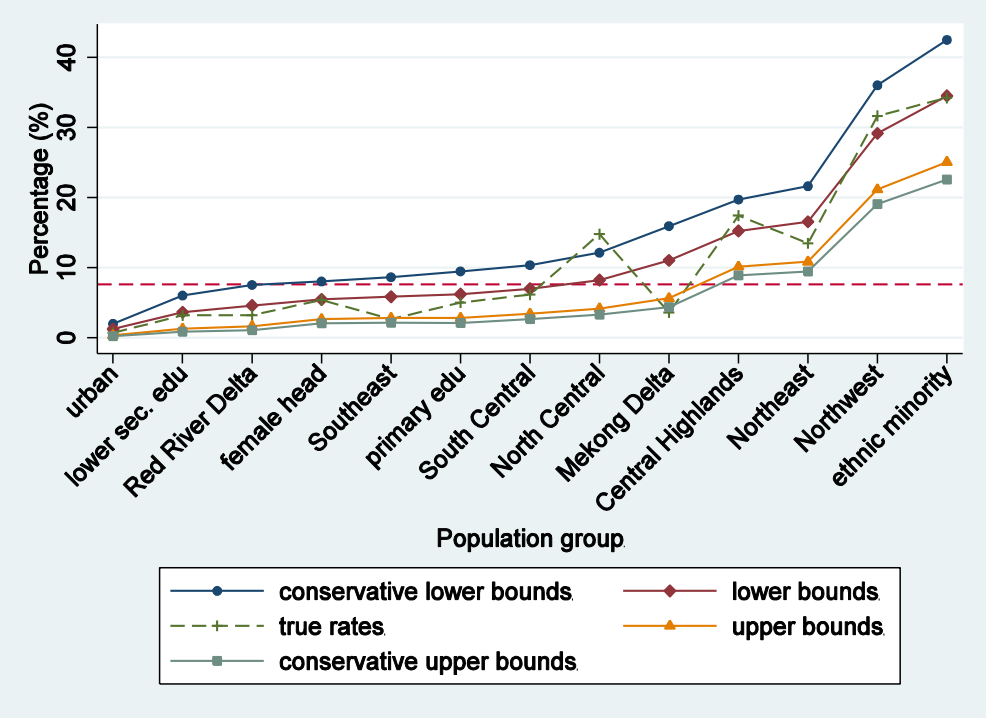


Figure 3: Profiles for Those Who Were Poor in the First Period but Non-poor in the Second Period, Vietnam 2006- 2008

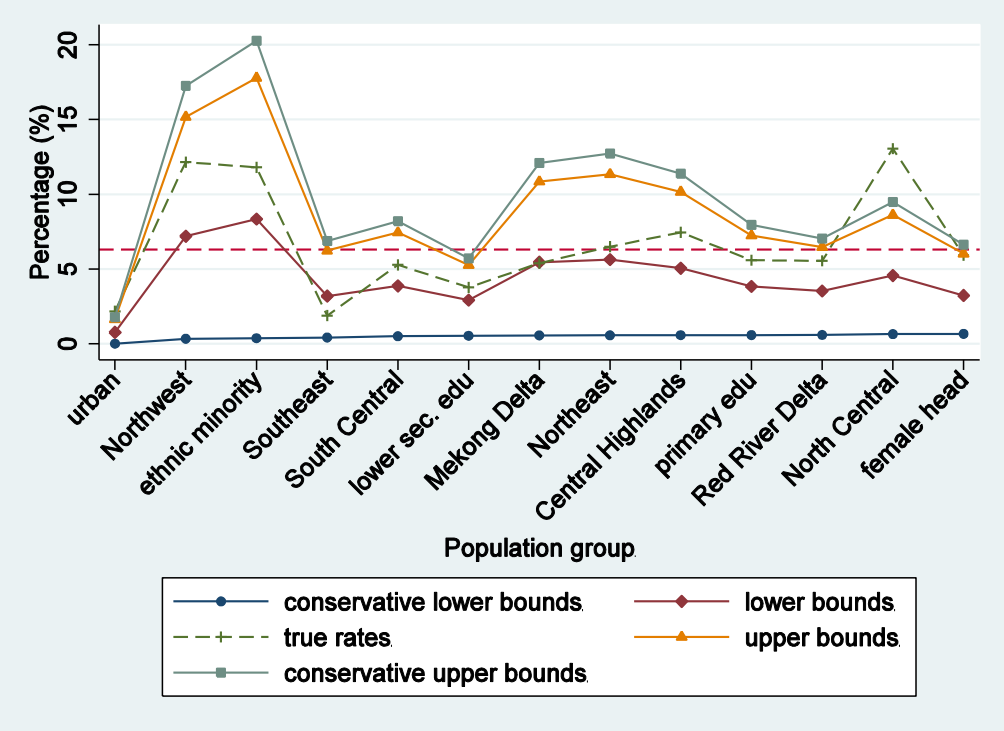


Figure 4: Profiles for Those Who Were Non-poor in the First Period but Poor in the Second Period, Vietnam 2006- 2008

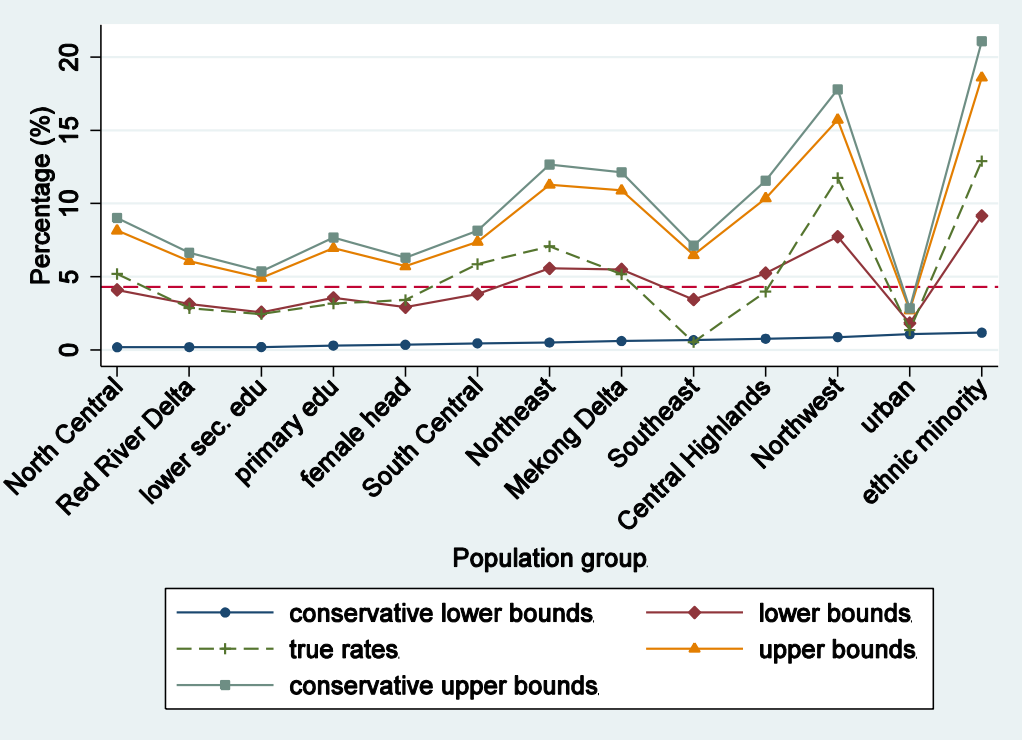
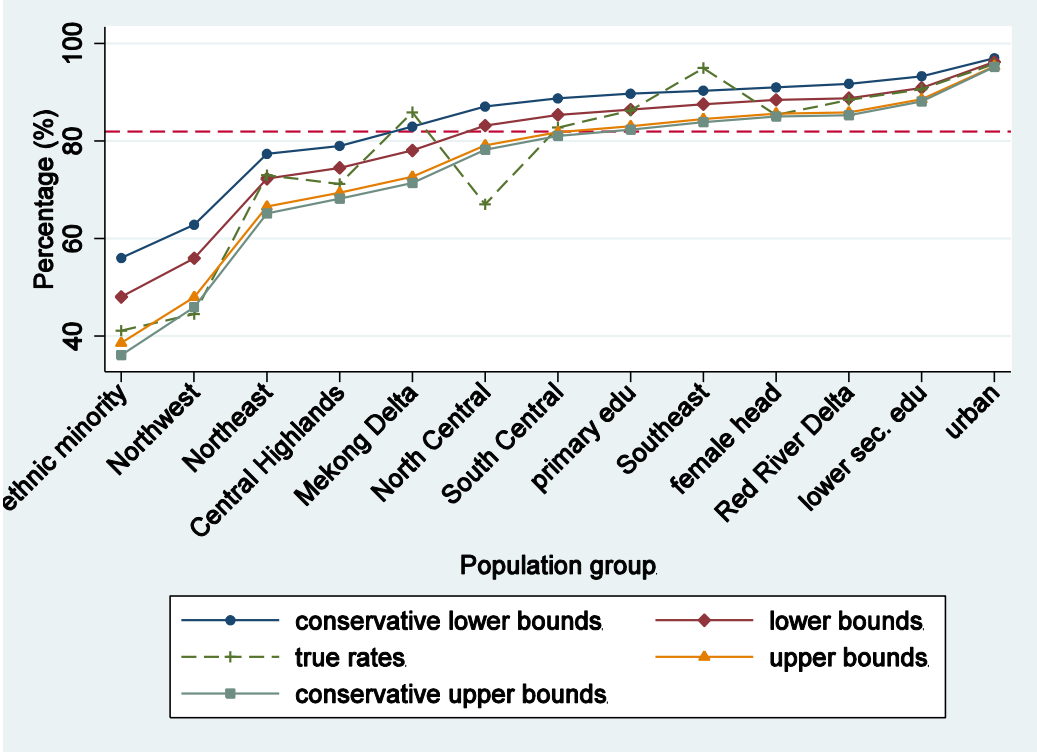


Figure 5: Profiles for Those Who Were Non-poor in Both Periods, Vietnam 2006- 2008



APPENDICES FOR ONLINE PUBLICATION ONLY

Appendix 1

Proof of Theorem 1 and Corollaries 1.1 and 1.2

The probability a household is poor in the first period but non-poor in the second period can be written as

$$\begin{aligned} P(y_{i1} < z_1 \bigcap y_{i2} > z_2) &= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1} \bigcap \varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \\ &= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2} \bigcap \varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \\ &= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2} \mid \varepsilon_{i1} < z_1 - \beta_1' x_{i2}) \end{aligned} \quad \text{(A1.1a)}$$

where the second line follows from replacing x_{i1} with x_{i2} by Assumption 1[27], and the third line follows from the multiplication rule for conditional probabilities.[28] Since the probability $P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2} \mid \varepsilon_{i1} \geq z_1 - \beta_1' x_{i2})$ (*) is non-negative by definition, we then have

$$\begin{aligned} P(y_{i1} < z_1 \bigcap y_{i2} > z_2) &\leq P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2} \mid \varepsilon_{i1} < z_1 - \beta_1' x_{i2}) \\ &\quad + P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2} \mid \varepsilon_{i1} \geq z_1 - \beta_1' x_{i2}) \\ &= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \end{aligned} \quad \text{(A1.2)}$$

where the second line follows from the partition rule.[29]

Our upper bound estimate of mobility can be written as

$$P(y_{i1}^{2U} < z_1 \bigcap y_{i2} > z_2) = P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \quad \text{(A1.3)}$$

where the right-hand side results when the two error terms ε_{i1} and ε_{i2} are completely independent of each other.

Thus combining (A1.2) and (A1.3) it follows that

$$P(y_{i1}^{2U} < z_1 \bigcap y_{i2} > z_2) \geq P(y_{i1} < z_1 \bigcap y_{i2} > z_2) \quad \text{(A.1.4)}$$

which establishes the upper bound estimate of mobility. Incidentally, the probability (*) is the bias for the upper bound estimate of mobility, which establishes Corollary 1.1.

Then subtracting each of the terms in (A1.4) from $P(y_{i2} > z_2)$, we would have

---

[27] Note that we can directly replace x_{i1} with x_{i2} if x contains only time-invariant variables. If x also contains deterministic variables, then we would replace x_{i1} with the period 1 values determined by knowing x_{i2}. We abstract from this case to simplify notation, since the key idea remains the same.

[28] Strictly speaking, we need $P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) > 0$ to derive the third line, which is satisfied as long as the poverty rate is not zero for period 1. Also note that the equality signs "=" in all the equal-or-greater-than "≥" signs inside parentheses for the following probabilities are optional since household consumptions (and their error terms) are continuous variables.

[29] See, for example, Theorem 1.2.11 in Casella and Berger (2002).

$$P(y_{i2} > z_2) - P(y_{i1}^{2U} < z_1 \bigcap y_{i2} > z_2) \le P(y_{i2} > z_2) - P(y_{i1} < z_1 \bigcap y_{i2} > z_2)$$

or equivalently, using the partition rule again,

$$P(y_{i1}^{2U} \ge z_1 \bigcap y_{i2} > z_2) \le P(y_{i1} \ge z_1 \bigcap y_{i2} > z_2) \quad \text{(A1.5)}$$

which establishes Corollary 1.2. And it is rather straightforward to show the remaining cases.

Proof of Theorem 2 and Corollaries 2.1 and 2.2

The probability a household is poor in the first period but non-poor in the second period in (A1.1a) can also be rewritten as

$$\begin{aligned}
&P(y_{i1} < z_1 \bigcap y_{i2} > z_2) = P(\beta_1' x_{i1} + \varepsilon_{i1} < z_1 \bigcap \beta_2' x_{i2} + \varepsilon_{i2} > z_2) \\
&= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1}) + P(\varepsilon_{i2} > z_2 - \beta_2' x_{i2}) - P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1} \bigcup \varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \\
&= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1}) + \left[1 - P(\varepsilon_{i2} \le z_2 - \beta_2' x_{i2})\right] - P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1} \bigcup \varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \quad \text{(A1.1b)} \\
&= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1}) - P(\varepsilon_{i2} \le z_2 - \beta_2' x_{i2}) + \left[1 - P(\varepsilon_{i1} < z_1 - \beta_1' x_{i1} \bigcup \varepsilon_{i2} > z_2 - \beta_2' x_{i2})\right] \\
&= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) - P(\varepsilon_{i2} \le z_2 - \beta_2' x_{i2}) + \left[1 - P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2} \bigcup \varepsilon_{i2} > z_2 - \beta_2' x_{i2})\right]
\end{aligned}$$

where the second and third lines follow from the basic properties of probability, [30] the fourth line follows from rearranging expressions, and the fifth line follows from replacing x_{i1} with x_{i2} using Assumption 1.

Our lower bound estimate of mobility is

$$\begin{aligned}
&P(y_{i1}^{2L} < z_1 \bigcap y_{i2} > z_2) = P(\varepsilon_{i2} < z_1 - \beta_1' x_{i2} \bigcap \varepsilon_{i2} > z_2 - \beta_2' x_{i2}) \\
&= P(\varepsilon_{i2} < z_1 - \beta_1' x_{i2}) - P(\varepsilon_{i2} \le z_2 - \beta_2' x_{i2}) \quad \text{(A1.6)} \\
&= P(\varepsilon_{i1} < z_1 - \beta_1' x_{i2}) - P(\varepsilon_{i2} \le z_2 - \beta_2' x_{i2})
\end{aligned}$$

where the last line follows when ε_{i1} has perfect correlation with ε_{i2}. Since the third term on the right-hand side in the last line in equation (A1.1b) is non-negative by definition, combining (A1.1b) and (A1.6) it follows that $P(y_{i1}^{2L} < z_1 \bigcap y_{i2} > z_2) \le P(y_{i1} < z_1 \bigcap y_{i2} > z_2)$ (A1.7)

which establishes our conservative lower bound of mobility. Incidentally, the third term on the right-hand side in the last line in equation (A1.1b) is the bias for the lower bound estimate of mobility, which establishes Corollary 2.1.

Then subtracting each of the terms in (A1.7) from $P(y_{i2} > z_2)$, we would have

$$P(y_{i2} > z_2) - P(y_{i1}^{2L} < z_1 \bigcap y_{i2} > z_2) \ge P(y_{i2} > z_2) - P(y_{i1} < z_1 \bigcap y_{i2} > z_2)$$

or equivalently

[30] See, for example, Theorem 1.2.9 in Casella and Berger (2002).

$$P(y_{i1}^{2L} \geq z_1 \bigcap y_{i2} > z_2) \geq P(y_{i1} \geq z_1 \bigcap y_{i2} > z_2) \quad \text{(A1.8)}$$

which establishes Corollary 2.2. And it is rather straightforward to show the remaining cases.

Proof of Theorem 3

When at least one independent variable is measured with error, the vector of household *i*'s true variables x_{ij}^* for j= 1, 2, are not observed, but instead we observe x_{ij} that are measured with errors. Similarly, if there are measurement errors in household consumption, true household consumption y_{ij}^* is not measured, but we only observe y_{ij}. The linear projection of true household consumption on true household characteristics in period j in equations (1) and (2) then becomes

$$y_{ij}^* = \beta_j' x_{ij}^* + u_{ij} \quad \text{(A1.9)}$$

The true and observed variables are postulated to have the following relationship

$$x_{ij} = x_{ij}^* + \tau_{ij} \quad \text{(A1.10)}$$

$$y_{ij} = y_{ij}^* + \upsilon_{ij} \quad \text{(A1.11)}$$

where τ_{ij} and υ_{ij} are the measurement errors. In the classical measurement error model, τ_{ij} and υ_{ij} are assumed to be uncorrelated respectively with the true variables x_{ij}^* and y_{ij}^*, as well as both uncorrelated with the model error u_{ij}. In the non-classical error model, there is less restriction on the correlation between these measurement errors and the true variables and τ_{ij} and υ_{ij} can be assumed to be correlated with x_{ij}^* and y_{ij}^*.

However, regardless of the correlation between the measurement errors and the true variables, using equations (A1.10) and (A1.11), we can rewrite (A1.9) as

$$y_{ij} = \beta_j' x_{ij} + (u_{ij} - \beta_j' \tau_{ij} - \upsilon_{ij}) \quad \text{(A1.12a)}$$

or conveniently in a more general format

$$y_{ij} = \beta_j' x_{ij} + \varepsilon_{ij} \quad \text{(A1.12b)}$$

Equation (A1.12b) is identical to our original equations (1) and (2), which shows that measurement errors do not affect our results in the proofs for Theorems 1 and 2. Indeed, equations (1) and (2) only provide the linear projection of observed household consumption on observed household characteristics, where we make no assumption about the correlation between the measurement errors and the true variables, except that they do not cause the autocorrelation of the ε_{ij} to become negative. Thus, the lower bound (which is based only on assuming the autocorrelation is less than or equal to one) will continue to be a lower bound,

while the upper bound will still be an upper bound with classical measurement error (since this will not change the autocorrelation of the ε_{ij} term), and will be an upper bound with non-classical measurement error provided this non-classical error doesn't induce negative autocorrelation. This could be violated if the measurement error in consumption is strongly negatively autocorrelated enough to offset the positive autocorrelation in the genuine consumption residual, which doesn't seem that likely in practice as evidenced by the positive overall autocorrelations of the ε_{ij} seen in our empirical applications.

Appendix 2

Figure 2.1: Distribution Graphs for the Residuals, Indonesia and Vietnam

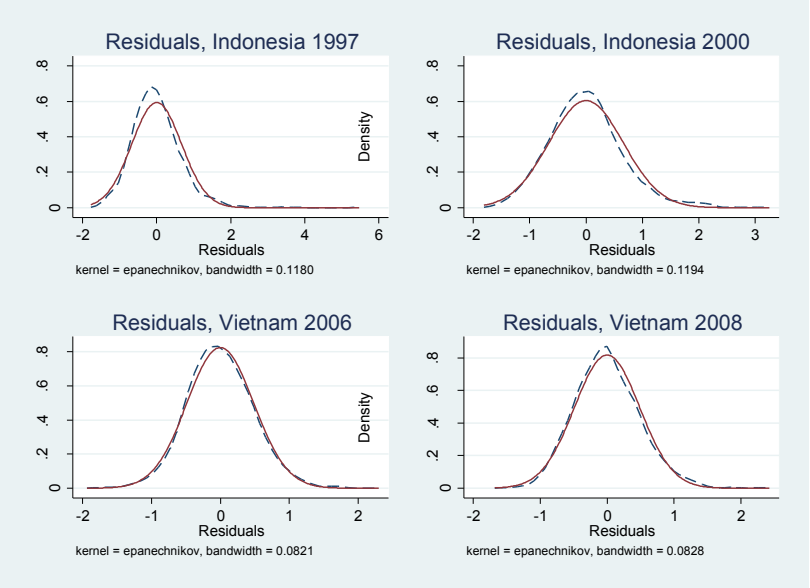


Table 2.1a: Estimated Parameters of Household Consumption, Vietnam 2006

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Heads' age	0.012***	0.010***	0.009***	0.009***	0.010***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Head is female	0.118***	0.009	0.030	0.023	-0.071**	-0.029
	(0.037)	(0.036)	(0.035)	(0.035)	(0.034)	(0.028)
Head's years of schooling	0.064***	0.057***	0.047***	0.046***	0.042***	0.021***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Ethnic majority groups	0.437***	0.333***	0.272***	0.254***	0.224***	0.194***
	(0.038)	(0.047)	(0.042)	(0.042)	(0.039)	(0.035)
Urban in 2006		0.297***	0.285***	0.215***	0.201***	0.088**
		(0.041)	(0.039)	(0.040)	(0.040)	(0.036)
Poor as classified by government in 2006			-0.435***	-0.434***	-0.417***	-0.238***
			(0.034)	(0.034)	(0.031)	(0.030)
Head works in agriculture only				0.070**	0.056**	0.038*
				(0.027)	(0.026)	(0.022)
Head works in wage only				0.197***	0.191***	0.099***
				(0.042)	(0.040)	(0.033)
Head works in service only				0.187***	0.192***	0.049
				(0.042)	(0.040)	(0.035)
Household size					-0.080***	-0.102***
					(0.009)	(0.008)
Number of children age 0 to 5					-0.068***	-0.062***
					(0.021)	(0.017)
Household owns a tivi						0.153***
						(0.032)
Household owns a motobicycle						0.283***
						(0.023)
Household owns a refrigerator						0.229***
						(0.032)
Household owns a wasing machine						0.172***
						(0.055)
Household owns an air conditioner						0.417***
						(0.109)
Household owns toilet						0.152***
						(0.043)
Drinking water from own running water or bottled water						0.034
						(0.039)
Constant	7.057***	7.601***	7.849***	7.791***	8.178***	7.926***
	(0.090)	(0.147)	(0.135)	(0.130)	(0.134)	(0.112)
Adjusted R2	0.334	0.494	0.548	0.559	0.600	0.710
σ	0.500	0.436	0.412	0.407	0.387	0.330
N	1334	1334	1334	1334	1334	1334

Note: 1. *p<0 .1, **p<0.05, ***p<0.01; robust standard errors in parentheses accounts for clustering at the primary sampling unit level.

2. Models 2 to 6 control for province dummy variables.

3. All estimates are obtained using cross sectional data.

Table 2.1b: Estimated Parameters of Household Consumption, Indonesia 1997

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Heads' age	0.007***	0.007***	0.007***	0.006***	0.007***	0.004**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Head is female	0.152***	0.142**	0.154***	0.209***	-0.013	-0.003
	(0.058)	(0.056)	(0.057)	(0.062)	(0.057)	(0.053)
Head's years of schooling	0.052***	0.053***	0.052***	0.045***	0.046***	0.026***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Head's birth place is small town	0.093**	0.087*	0.069	0.062	0.046	0.015
	(0.046)	(0.050)	(0.050)	(0.050)	(0.048)	(0.042)
Head's birth place is big city	0.092	0.045	0.038	0.042	0.054	0.015
	(0.082)	(0.086)	(0.087)	(0.084)	(0.079)	(0.073)
Head's birth place is other	-0.076	-0.091	-0.114	-0.072	-0.392	-0.460
	(0.424)	(0.432)	(0.433)	(0.449)	(0.397)	(0.422)
Urban		0.015	-0.006	-0.026	0.014	-0.094*
		(0.045)	(0.051)	(0.054)	(0.052)	(0.051)
Community rate of electrification			0.002**	0.002**	0.003***	0.002**
			(0.001)	(0.001)	(0.001)	(0.001)
Community has a primary school			0.077	0.058	0.093	0.099
			(0.088)	(0.084)	(0.081)	(0.075)
Head is self-employed				0.312***	0.269***	0.251***
				(0.084)	(0.073)	(0.063)
Head works for the government				0.475***	0.411***	0.289***
				(0.103)	(0.095)	(0.084)
Head works in the private sector				0.199**	0.146*	0.154**
				(0.088)	(0.078)	(0.069)
Head is unpaid family worker				0.476*	0.450*	0.382*
				(0.280)	(0.263)	(0.218)
Household farms				-0.102**	-0.067	-0.023
				(0.050)	(0.046)	(0.042)
Household size					-0.311***	-0.345***
					(0.040)	(0.039)
Household size squared					0.019***	0.021***
					(0.003)	(0.003)
Number of children age 0 to 5					-0.101***	-0.084***
					(0.025)	(0.023)
Log of housing floor space (m2)						0.117***
						(0.026)
Main drinking water from pipe						0.100**
						(0.040)
Household owns a tivi						0.188***
						(0.031)
Constant	11.642***	11.383***	11.184***	10.960***	11.999***	11.782***
	(0.123)	(0.154)	(0.178)	(0.208)	(0.208)	(0.312)
Adjusted R2	0.193	0.210	0.215	0.231	0.329	0.421
σ	0.678	0.670	0.668	0.662	0.618	0.574
N	1659	1659	1659	1659	1659	1659

Note: 1. *p<0 .1, **p<0.05, ***p<0.01; robust standard errors in parentheses accounts for clustering at the primary sampling unit level.

2. Models 2 to 6 include dummy variables for provinces, languages spoken at home, religions, education lev of head's father. Models 3 to 6 include dummy variables for community road types.

Models 6 includes dummy variables for types of cooking fuel and primary roof materials.

3. All estimates are obtained using cross sectional data.

Table 2.2: Estimated Parameters of Household Consumption Using Actual Panel Data for Different Countries

	Vietnam	Bosnia-Herzegovina	Lao PDR	Nepal	Peru
	2006-08	2001-04	2002/03-2007/08	1995/96- 2003/04	2004-06
Age	0.020***	0.010***	0.030***	0.030***	0.012***
	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)
Female	0.042*	0.233***	0.037	0.310***	0.184***
	(0.022)	(0.035)	(0.065)	(0.065)	(0.026)
Years of schooling	0.048***	0.037***	0.042***	0.065***	0.057***
	(0.002)	(0.004)	(0.003)	(0.007)	(0.003)
Ethnic majority groups/ upper caste	0.379***		0.145***	-0.104**	0.150***
	(0.023)		(0.025)	(0.049)	(0.023)
Bosniac		-0.123***			
		(0.041)			
Serb		-0.088**			
		(0.041)			
Urban	0.362***	-0.084***	0.131***	0.341***	0.440***
	(0.022)	(0.026)	(0.027)	(0.078)	(0.023)
Constant	6.939***	7.213***	10.470***	7.586***	4.062***
	(0.050)	(0.103)	(0.060)	(0.127)	(0.059)
σ_u	0.37	0.35	0.34	0.35	0.41
σ_v	0.29	0.40	0.42	0.43	0.35
ρ	0.62	0.43	0.40	0.39	0.58
R^2	0.37	0.07	0.15	0.27	0.40
Number of households	2728	1341	2000	419	2665
Total no of observations	5456	2682	3877	838	4095

Note: 1. *p<0 .1, **p<0.05, ***p<0.01; robust standard errors in parentheses accounts for clustering at the individual level.

2. Household heads' ages are restricted to between 25 and 55 in the first round.